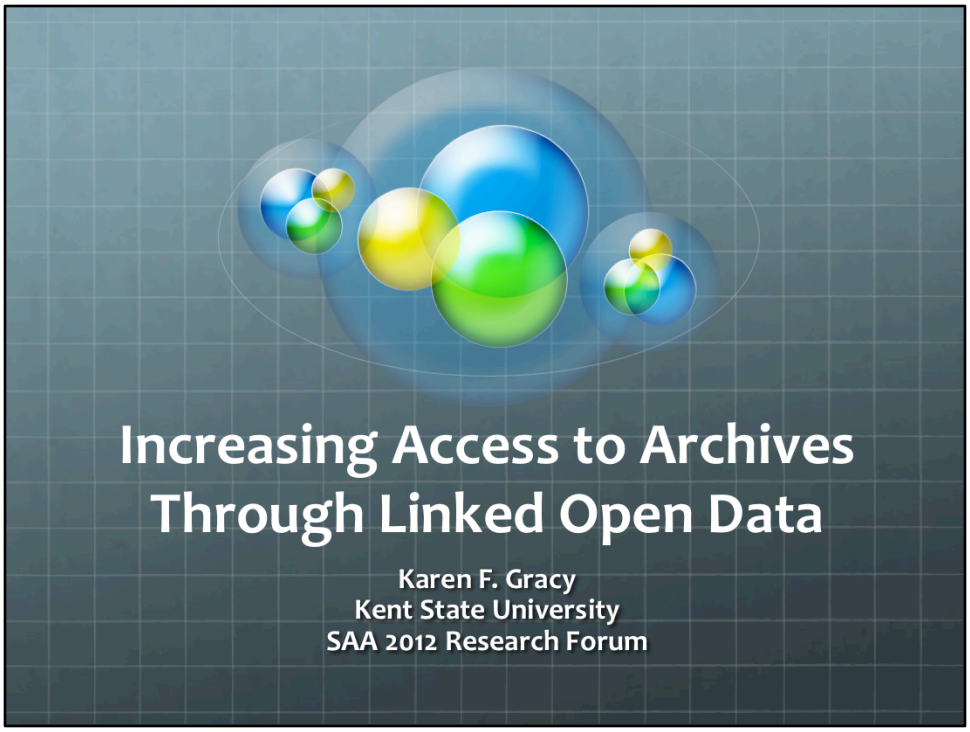


This presentation will report on a research study being conducted on how archives, libraries, and other cultural institutions can benefit from the data and metadata resources that have been created as a result of the Linked Open Data (LOD) movement. The LOD movement aims to connect related data formerly isolated in repositories (often called “silos”) and not previously linked. Archival institutions generate valuable data through tools such as finding aids, cataloging records, and databases, but this information is often sequestered inside archival information systems that cannot be easily linked to other data sources. Linked data technologies and LOD resources can enable archives to enhance their already existing finding aids and other services, allow archival users to obtain new information and knowledge relevant to materials in the archive, and make available archival collections to the users of external databases.

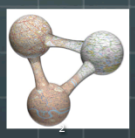
A Research Report on …

The Metadata Vocabulary Junction Project:

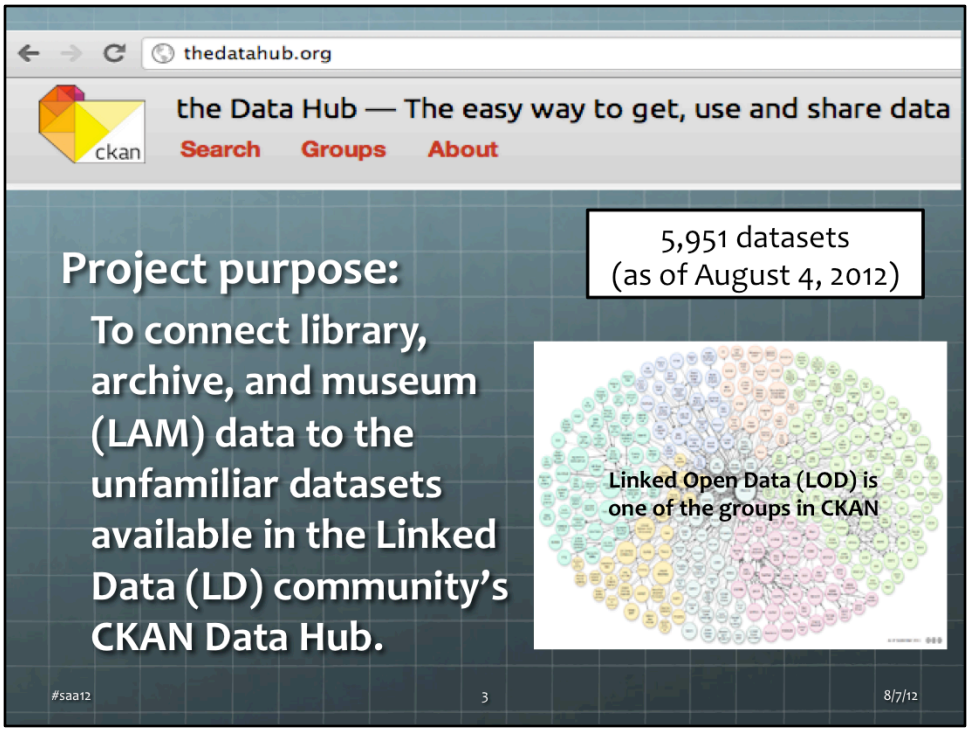
Connecting Library Data to the Unfamiliar Data and Metadata Resources in the LOD Universe

PI: Marcia Lei Zeng, Ph.D.
Co-PI: Karen Gracy, Ph.D.
Research Assistants:
Laurence Skirvin, MLIS; Sammy Davidson, MLIS; Shadi Shakeri, MLIS; Riley Stoermer, MLIS

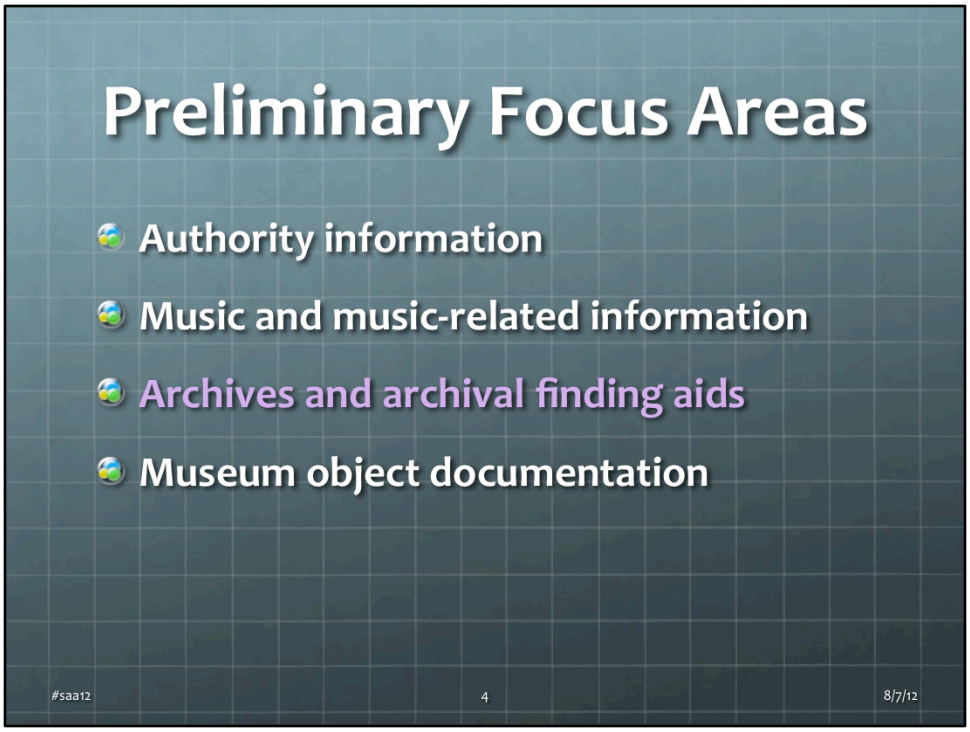
http://lod-lam.slis.kent.edu/



#saa12 2 8/7/12

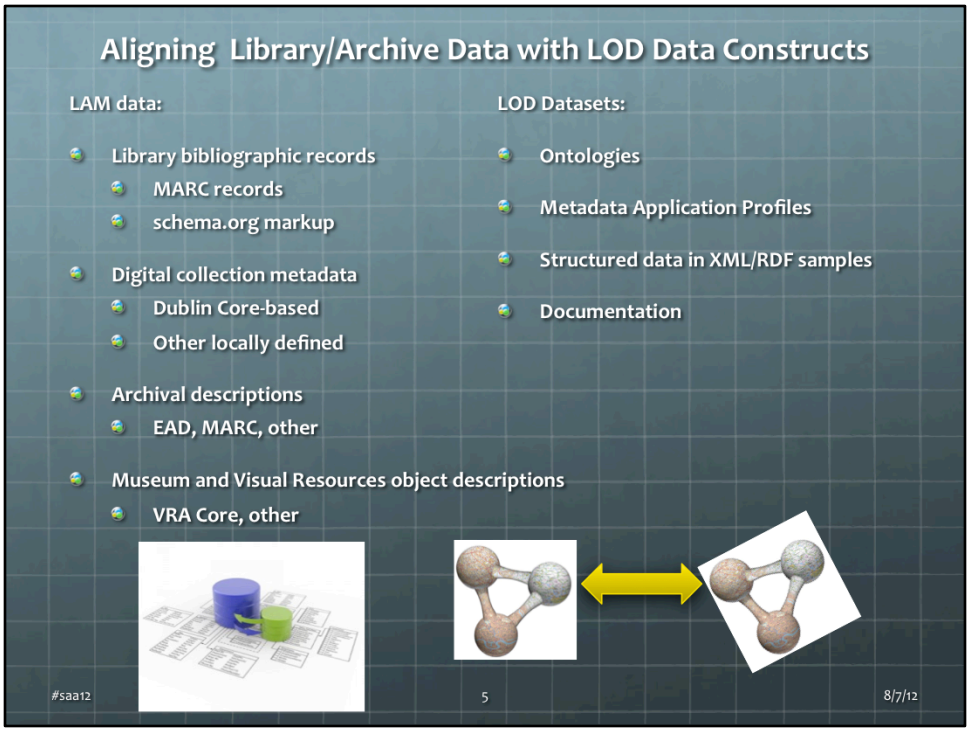

thedatahub.org
the Data Hub — The easy way to get, use and share data
ckan
Search
Groups
About
Project purpose:
To connect library, archive, and museum (LAM) data to the unfamiliar datasets available in the Linked Data (LD) community's CKAN Data Hub.
5,951 datasets
(as of August 4, 2012)
Linked Open Data (LOD) is one of the groups in CKAN
#saa12
3
8/7/12

This presentation will report on a research study being conducted on how archives, libraries, and other cultural institutions can benefit from the data and metadata resources that have been created as a result of the Linked Open Data (LOD) movement. The LOD movement aims to connect related data formerly isolated in repositories (often called "silos") and not previously linked. Archival institutions generate valuable data through tools such as finding aids, cataloging records, and databases, but this information is often sequestered inside archival information systems that cannot be easily linked to other data sources. Linked data technologies and LOD resources can enable archives to enhance their already existing finding aids and other services, allow archival users to obtain new information and knowledge relevant to materials in the archive, and make available archival collections to the users of external databases.



In this phase of the project, we are focusing on archives and archival finding aids. We have just finished a paper on music and music-related information, currently under consideration for publication right now.

At the end of this project, we'll be issuing a final report which will be made available on the project website.



Trying to sort out among the many diverse metadata vocabularies

Dozens of metadata standards were developed by various user communities in the 1990s in response to the rise of the World Wide Web and the invention of digital libraries. Examples of popular standards used in memory institutions include:

- Dublin Core Element Set (DC)
- Encoded Archival Description (EAD)
- Metadata Object Description Schema (MODS)
- Categories for the Description of Works of Art (CDWA)
- Core Categories for Visual Resources (VRA Core).

In the 2000s, more specialized metadata standards have joined the crowd through two main avenues:

(a) independently developed schemas such as PBCore for public broadcasting, the IPTC photo metadata standard that comes with Photoshop software, and XMP metadata for all Adobe documents;
(b) application profiles of an existing metadata standard, such as Dublin Core; often developed by institutions to expand the capabilities of the standard beyond the defined elements.

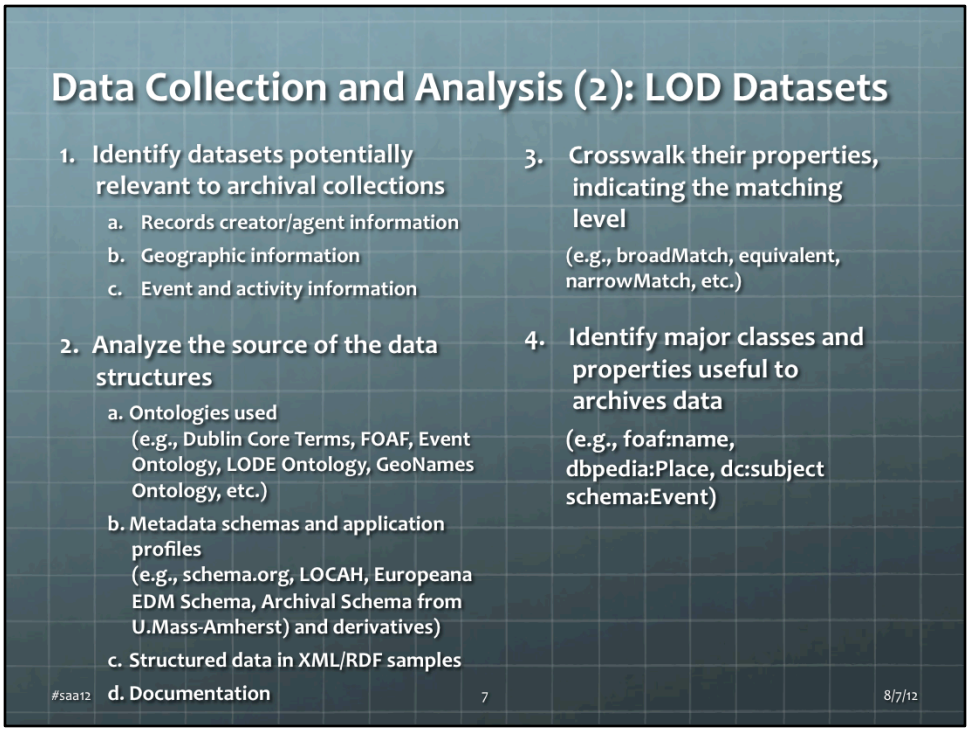
Data Collection and Analysis (1): Archives Information Systems

1. Analyze archival descriptive and authority standards to determine what information may be useful as linked data:
   - Encoded Archival Description
   - Encoded Archival Context
2. Review finding aid exemplars from several sources to discover how standards are applied:
   - OhioLink EAD Repository
   - Online Archive of California
   - Library of Congress Finding Aids
3. Identify and review other potential sources of information as needed

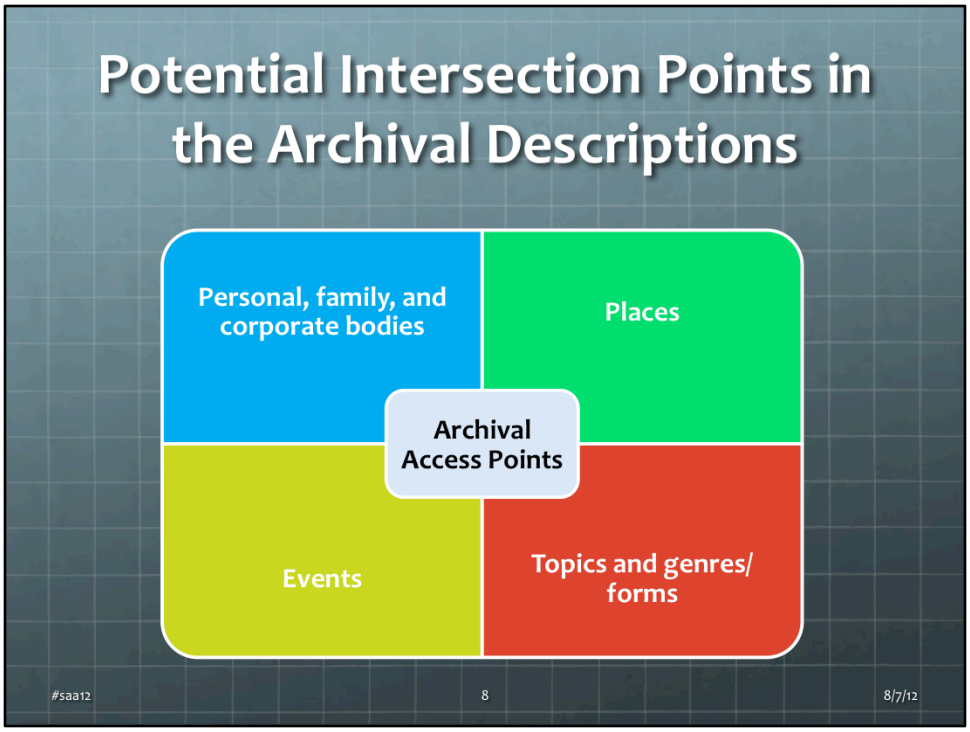
4. Then, break down the "records" and find:
   - Common elements
   - Possibilities for linking archival data elements to LOD classes and properties
   - Major access points
   - Hidden access points

#saa12 6 8/7/12

Our research process includes two major steps: for the first part, we have been examining archival descriptive and authority standards to see what information may be useful as linked data. We are also looking at finding aid and authority exemplars, which we pulled from major sources of archival finding aids such as the OhioLink EAD repository. After reviewing standards and sample finding aids and authority records, we identify common elements among them and consider the possibilities for linking archival information to linked open data properties—we are looking at both major access points and potential hidden access points not currently able to be linked easily because of how they are currently encoded.


Data Collection and Analysis (2): LOD Datasets
1. Identify datasets potentially relevant to archival collections
a. Records creator/agent information
b. Geographic information
c. Event and activity information
2. Analyze the source of the data structures
a. Ontologies used (e.g., Dublin Core Terms, FOAF, Event Ontology, LODE Ontology, GeoNames Ontology, etc.)
b. Metadata schemas and application profiles (e.g., schema.org, LOCAH, Europeana EDM Schema, Archival Schema from U.Mass-Amherst) and derivatives)
c. Structured data in XML/RDF samples
d. Documentation
3. Crosswalk their properties, indicating the matching level
(e.g., broadMatch, equivalent, narrowMatch, etc.)
4. Identify major classes and properties useful to archives data
(e.g., foaf:name, dbpedia:Place, dc:subject schema:Event)
#saa12
7
8/7/12

In the second part of the research, we have identified datasets that have potential to be relevant to users of archival collections. These datasets are analyzed to discover what ontologies, metadata schemas, and/or application profiles that are used. We also look at sample records and any documentation that we can find, using information found on the Data Hub. Then, we attempt to crosswalk across datasets, to see how well a particular type of information will match. We use categories common to ontology alignent such as equivalent, close match, broad match, narrow match, and related. Through this process, we identify the major classes and properties useful to archives data.

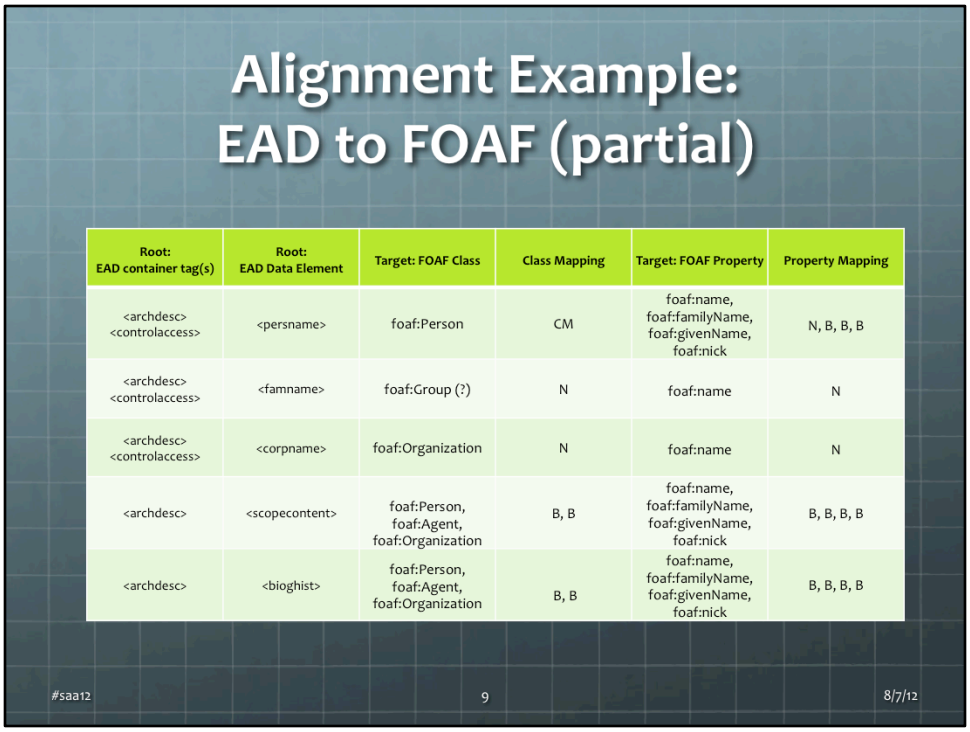


At this stage of our research, we are focusing on data elements, classes, and properties in four major categories:

1. Personal, family, and corporate bodies
2. Places
3. Events
4. Topics and genres/forms

We see the potential for linking from archival descriptions in these four types of things or entities to numerous data sets. Some examples of datasets that could prove helpful for archives users include:

1. GeoNames: The GeoNames geographical database is available for download free of charge under a creative commons attribution license. It contains over 10 million geographical names and consists of over 8 million unique features whereof 2.8 million populated places and 5.5 million alternate names. By linking geographic names found in archival descriptions to GeoNames, institutions could enrich finding aids and records with additional information about locations associated with collections and creators.
2. Census data: The United States Census data includes information on names, birth place, residence, and demographic data. By linking from archival descriptions to



This is one example of a mapping table that is generated as a result of the analysis we're doing with the standards, ontologies, and datasets. In this mapping of EAD to FOAF, which focuses on personal, family, and corporate body names, we can make a couple of general statements about how well these two data models relate. First, we can say that for the EAD data elements <persname>, <famname>, and <corpname>, at the class level they are narrower than the closest FOAF class equivalent, except in the case of Person. Matching to the property level, however, FOAF tends to be narrower in scope than the closest equivalent in EAD. Matching names found in EAD records to the property foaf:name would provide a basic level of interoperability, but would not differentiate among first, last, and nicknames. Thus, EAD and FOAF are not completely equal in the granularity of their data models.

Another observation that we have made in EAD is that there are several places in the EAD record where names can be found, but which are not and cannot be tagged as such, such as in the ScopeContent and BiogHist tags. The potential to link additional names to external datasets is therefore untapped at this time, but revisers of the standard may wish to consider adding such functionality to future versions of EAD.

For those of you who wish to follow the progress of our project, I direct you to our website found at the address listed on the slide: http://lod-lam.slis.kent.edu/ . Please feel free to contact me or my research partner, Marcia Zeng if you have any questions or would like to share information on a related project. We are eager to examine the data models of new archives linked data projects, and would love to speak with you about what we're doing.